注意に基づくニューラル機械翻訳への効果的なアプローチ

Minh-Thang Luong **Hieu Pham** **Christopher D. Manning**
Computer Science Department, Stanford University, Stanford, CA 94305
{lmthang,hyhieu,manning}@stanford.edu

Abstract

最近、翻訳時に原文の一部に選択的に焦点を当てることで、ニューラル機械翻訳(NMT)を改善するアテンションメカニズムが用いられている。しかし、注意に基づくNMTに有用なアーキテクチャを探求した研究はほとんどない。本論文では、注意メカニズムの2つのシンプルで効果的なクラス、すなわち、常にすべての原語に注目するグローバルアプローチと、一度に原語のサブセットだけを見るローカルアプローチについて検討する。英語とドイツ語のWMT翻訳タスクにおいて、両アプローチの有効性を両方向に実証する。ローカルアテンションでは、ドロップアウトのような既知の技術を既に取り入れた非アテンションシステムに対して、5.0 BLEUポイントの大幅な利得を達成する。異なる注意アーキテクチャを用いた我々のアンサンブルモデルは、WMT'15の英独翻訳タスクにおいて、25.9 BLEUポイントという新しい最先端の結果をもたらし、NMTとn-gramリランカーに裏打ちされた既存の最良システムよりも1.0 BLEUポイント改善された[1]。

1 Introduction

ニューラル機械翻訳(NMT)は、英語からフランス語(Luong et al., 2015)、英語からドイツ語(Jean et al., 2015)などの大規模翻訳タスクにおいて、最先端の性能を達成した。NMTは最小限のドメイン知識で済み、概念的に単純であるため、魅力的である。Luongら(2015)のモデルは、文末記号<eos>に達するまで、すべての原語を読み取る。

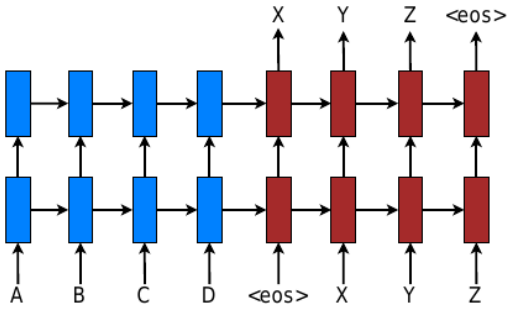

図1:ニューラル機械翻訳 – ソースシーケンスA B C Dをターゲットシーケンス X Y Zに変換するためのスタッキングリカレントアーキテクチャ。

その後、図1に示すように、一度に1つのターゲット単語を放出し始める。NMTは多くの場合、エンドツーエンドで学習される大規模なニューラルネットワークであり、非常に長い単語列に対してうまく汎化する能力を持つ。つまり、標準的なMTの場合のように、巨大なフレーズテーブルや言語モデルを明示的に格納する必要がないため、NMTのメモリフットプリントは小さい。最後に、NMTデコーダの実装は、標準的なMTの非常に複雑なデコーダとは異なり、簡単である(Koehn et al., 2003)。

並行して、「注意」の概念はニューラルネットワークの学習において最近人気を博しており、モデルは異なるモダリティ間、例えば動的制御問題における画像オブジェクトとエージェントアクション間(Mnih et al., 2014)、音声認識タスクにおける音声フレームとテキスト間(?)、あるいは画像キャプション生成タスクにおける画像とそのテキスト記述の視覚的特徴間(Xu et al., 2015)のアライメントを学習することができる。NMTの文脈では、Bahdanauら(2015)はこのような注意メカニズムを単語の共同翻訳と整列に適用することに成功している。我々の知る限り、NMTのための注意ベースのアーキテクチャの使用を探求する研究は他にない。

本研究では、シンプルさと有効性を念頭に置き、

[1] All our code and models are publicly available at http://nlp.stanford.edu/projects/nmt.

2つの新しいタイプの注意ベースのモデルを設計する:すべてのソース単語がアテンションされるグローバルなアプローチと、ソース単語のサブセットのみが一度に考慮されるローカルなアプローチである。前者のアプローチは(Bahdanau et al., 2015)のモデルに似ているが、アーキテクチャ的にはより単純である。後者は、(Xu et al., 2015)で提案されたハードアテンションモデルとソフトアテンションモデルの興味深いブレンドと見なすことができる:グローバルモデルやソフトアテンションよりも計算コストが低い。同時に、ハードアテンションとは異なり、ローカルアテンションはほとんどどこでも微分可能であるため、実装と訓練が容易である。[2] さらに、我々のアテンションベースモデルの様々なアライメント関数も検証する。

実験的に、我々のアプローチの両方が、英語とドイツ語の間のWMT翻訳タスクにおいて、両方向で効果的であることを実証する。我々の注意モデルは、脱落のような既知の技術をすでに取り入れている非注意システムに対して、最大5.0 BLEUのブーストをもたらす。英語からドイツ語への翻訳において、WMT'14とWMT'15の両方で、NMTモデルとn-gram LMリランカーに裏打ちされた従来のSOTAシステムを1.0 BLEU以上上回る、新しい最先端(SOTA)の結果を達成した。我々は、学習、長い文章を扱う能力、注意アーキテクチャの選択、アライメント品質、翻訳出力の観点から、我々のモデルを評価するために広範な分析を行う。

2 ニューラル機械翻訳

ニューラル機械翻訳システムは、原文 $x_1, \ldots, x_n$ を目的文 $y_1, \ldots, y_m$.[3] NMTの基本形は2つの要素からなる:(a)各原文の表現sを計算するエンコーダと、(b)一度に1つの目的語を生成し、したがって条件付き確率を次のように分解するデコーダである:

$$\log p(y|x) = \sum_{j=1}^{m} \log p\left(y_j|y_{<j}, \boldsymbol{s}\right) \quad (1)$$

A natural choice to model such a decomposition in the decoder is to use a

[2]There is a recent work by Gregor et al. (2015), which is very similar to our local attention and applied to the image generation task. However, as we detail later, our model is much simpler and can achieve good performance for NMT.

[3]All sentences are assumed to terminate with a special "end-of-sentence" token <eos>.

リカレントニューラルネットワーク(RNN)アーキテクチャは、(Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Cho et al., 2014; Bahdanau et al., 2015; Luong et al., 2015; Jean et al., 2015)などの最近のNMT作業のほとんどが共通している。しかし、どのRNNアーキテクチャをデコーダに使用し、エンコーダがどのように原文表現s, を計算するかという点では異なる。

Kalchbrenner and Blunsom (2013)は、デコーダに標準的な隠れユニットを持つRNNを用い、原文表現の符号化に畳み込みニューラルネットワークを用いた。一方、Sutskeverら(2014)とLuongら(2015)は、エンコーダとデコーダの両方にLong Short-Term Memory(LSTM)隠れユニットを持つRNNを複数層積み重ねた。Choら(2014)、Bahdanauら(2015)、Jeanら(2015)はいずれも、両コンポーネントにLSTMにインスパイアされた隠れユニットであるゲーテッドリカレントユニット(GRU)を持つRNNの異なるバージョンを採用している[4]。

より詳細には、各単語y_jの復号確率を次のようにパラメータ化できる:

$$p\left(y_j|y_{<j}, \boldsymbol{s}\right) = \text{softmax}\left(g\left(\boldsymbol{h}_j\right)\right) \quad (2)$$

ここで、$\boldsymbol{h}_j$はRNN隠れユニットであり、抽象的に次のように計算される:

$$\boldsymbol{h}_j = f(\boldsymbol{h}_{j-1}, \boldsymbol{s}), \quad (3)$$

ここで、fは前の隠れ状態が与えられたときに現在の隠れ状態を計算し、バニラRNNユニット、GRU、LSTMユニットのいずれかである。(Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Cho et al., 2014; Luong et al., 2015)では、ソース表現sはデコーダの隠れ状態を初期化するために一度だけ使われる。一方、(Bahdanau et al., 2015; Jean et al., 2015)やこの研究では、s, は、実際には、翻訳プロセスの全過程を通して参照されるソース隠れ状態のセットを意味する。このようなアプローチは注意メカニズムと呼ばれ、次に説明する。

本研究では、(Sutskever et al., 2014; Luong et al., 2015)に従い、図1に示すように、NMTシステムにスタッキングLSTMアーキテクチャを使用する。

[4] 後者の2つの作品はエンコーダに双方向RNNを利用した以外は、すべて単一のRNN層を使用した。

[5] (Bahdanau et al., 2015)のように、現在予測されている単語y_jのような他の入力でgを提供できる。

我々は(Zaremba et al., 2015)で定義されたLSTMユニットを使用する。我々の学習目的は以下のように定式化される：

$$J_t = \sum_{(x,y)\in\mathbb{D}} -\log p(y|x) \qquad (4)$$

Dを並列学習コーパスとする。

3 注意に基づくモデル

我々の様々な注意に基づくモデルは、大域的と局所的の2つに大別される。これらのクラスは、「注意」がすべてのソース位置に配置されるか、いくつかのソース位置にのみ配置されるかという点で異なっている。これら2つのモデルタイプをそれぞれ図2と図3に示す。

これら2種類のモデルに共通するのは、復号化段階の各時間ステップtにおいて、両アプローチとも、まずスタッキングLSTMの最上層の隠れ状態$\boldsymbol{h}_t$を入力とするという事実である。そして、現在のターゲット単語y_tを予測するのに役立つ、関連するソース側の情報を捕らえるコンテキストベクトル$\boldsymbol{c}_t$を導出する。これらのモデルは文脈ベクトル$\boldsymbol{c}_t$の導出方法が異なるが、その後のステップは同じである。

具体的には、ターゲット隠れ状態$\boldsymbol{h}_t$とソース側のコンテキストベクトル$\boldsymbol{c}_t$が与えられたとき、単純な連結層を採用して両ベクトルからの情報を結合し、以下のように注意隠れ状態を生成する：

$$\tilde{\boldsymbol{h}}_t = \tanh(\boldsymbol{W_c}[\boldsymbol{c}_t; \boldsymbol{h}_t]) \qquad (5)$$

次に、注意ベクトル$\boldsymbol{h}_t$はソフトマックス層を通して、次のように定式化された予測分布を生成する：

$$p(y_t|y_{<t}, x) = \text{softmax}(\boldsymbol{W_s}\tilde{\boldsymbol{h}}_t) \qquad (6)$$

次に、各モデルタイプがソース側のコンテキストベクトル$\boldsymbol{c}_t$を計算する方法を詳述する。

3.1 グローバルな注意

グローバル注意モデルのアイデアは、文脈ベクトル$\boldsymbol{c}_t$を導出する際に、エンコーダの全ての隠れ状態を考慮することである。このモデルタイプでは、現在のターゲット隠れ状態$\boldsymbol{h}_t$と各ソース隠れ状態$\bar{\boldsymbol{h}}_s$を比較することで、ソース側の時間ステップ数に等しいサイズを持つ可変長アライメントベクトル$\boldsymbol{a}_t$が導出される：

$$\begin{aligned}\boldsymbol{a}_t(s) &= \text{align}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) \qquad (7)\\ &= \frac{\exp\left(\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s)\right)}{\sum_{s'}\exp\left(\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_{s'})\right)}\end{aligned}$$

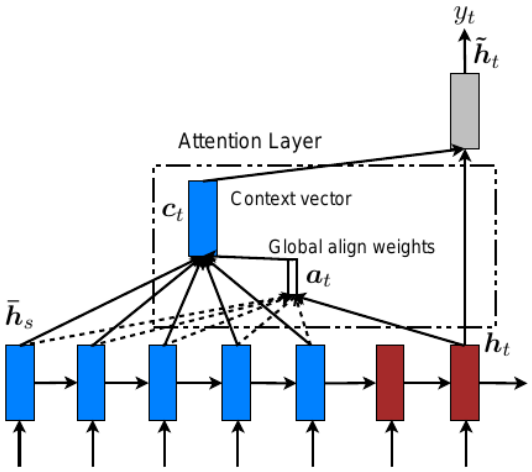

図2：グローバル注意モデル - 各時間ステップtにおいて、モデルは現在の目標状態$\boldsymbol{h}_t$と全てのソース状態$\bar{\boldsymbol{h}}_s$に基づいて可変長アライメント重みベクトル$\boldsymbol{a}_t$を推論する。グローバルコンテキストベクトル$\boldsymbol{c}_t$は、$\boldsymbol{a}_t$に従って、全てのソース状態に対する加重平均として計算される。

ここで、スコアはコンテンツベースの関数と呼ばれ、3つの異なる選択肢を検討する：

$$\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \begin{cases}\boldsymbol{h}_t^\top \bar{\boldsymbol{h}}_s & \text{dot}\\ \boldsymbol{h}_t^\top \boldsymbol{W_a}\bar{\boldsymbol{h}}_s & \text{general}\\ \boldsymbol{v}_a^\top \tanh\left(\boldsymbol{W_a}[\boldsymbol{h}_t; \bar{\boldsymbol{h}}_s]\right) & \text{concat}\end{cases}$$

また、注意に基づくモデルを構築する初期の試みでは、以下のように、ターゲット隠れ状態$\boldsymbol{h}_t$のみからアライメントスコアを計算する位置情報関数を使用する：

$$\boldsymbol{a}_t = \text{softmax}(\boldsymbol{W_a}\boldsymbol{h}_t) \qquad \text{location} \qquad (8)$$

アライメントベクトルを重みとして与え、コンテキストベクトル$\boldsymbol{c}_t$は全てのソース隠れ状態に対する加重平均として計算される[6]。

(Bahdanau et al., 2015)との比較 - 我々のグローバルアテンションアプローチはBahdanau et al. (2015)によって提案されたモデルと精神的に似ているが、オリジナルモデルから簡略化と一般化の両方を反映したいくつかの重要な違いがある。まず、図2に示すように、エンコーダとデコーダの両方で、単純に最上位LSTM層の隠れ状態を使用する。

[6] 式(8)は、すべての$\boldsymbol{a}_t$が同じ長さであることを意味する。短い文では$\boldsymbol{a}_t$の上部のみを使用し、長い文では末尾近くの単語を無視する。

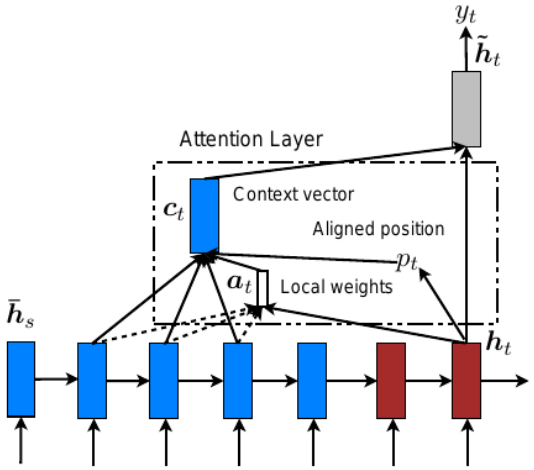

図3:局所注意モデル – このモデルはまず、現在のターゲット単語に対して1つの整列位置p_tを予測する。次に、ソース位置p_tを中心とするウィンドウを使用して、ウィンドウ内のソース隠れ状態の加重平均であるコンテキストベクトルc_tを計算する。重みa_tは、ウィンドウ内の現在の目標状態h_tとそれらのソース状態$\bar{h}_s$から推論される。

一方、Bahdanauら(2015)は、双方向エンコーダの前方および後方のソース隠れ状態と、彼らの非積層一方向デコーダのターゲット隠れ状態の連結を使用する。次に、$h_t \rightarrow a_t \rightarrow c_t \rightarrow h_t$ から、式(5)、式(6)、図2に詳述するように予測を行う。一方、任意の時刻tにおいて、Bahdanauら(2015)は前の隠れ状態$h_{t-1} \rightarrow a_t \rightarrow c_t \rightarrow h_t$ から構築し、その隠れ状態は予測を行う前に深い出力層と最大出力層を通過する。他の選択肢の方が優れていることを後で示す。

3.2 局所的注意

グローバルアテンションには、各ターゲットワードに対してソース側の全てのワードに注意を向けなければならないという欠点があり、これは高価であり、段落や文書など、より長いシーケンスを翻訳することを非現実的にする可能性がある。この欠点に対処するために、我々は、ターゲット単語ごとのソース位置の小さなサブセットのみに焦点を当てることを選択する局所的な注意メカニズムを提案する。

このモデルは、画像キャプション生成タスクに取り組むためにXuら(2015)によって提案されたソフト注意モデルとハード注意モデルの間のトレードオフからインスピレーションを得ている。

[7]We will refer to this difference again in Section 3.3.

彼らの研究において、ソフトアテンションとは、ソース画像内の全てのパッチに重みを「ソフト」に配置するグローバルアテンションアプローチを指す。一方、ハードアテンションは、一度にアテンションする画像のパッチを1つ選択する。推論時には安価であるが、ハードアテンションモデルは微分不可能であり、学習には分散削減や強化学習など、より複雑な技術が必要である。

我々の局所的注意メカニズムは、選択的にコンテキストの小さなウィンドウに焦点を当て、微分可能である。このアプローチは、ソフトアテンションで発生する高価な計算を回避すると同時に、ハードアテンションアプローチよりも学習が容易であるという利点がある。具体的には、まず時刻tにおける各ターゲット単語の位置p_tをモデルが生成する。次に、コンテキストベクトルc_tは、ウィンドウ内のソース隠れ状態の集合$[p_t - D, p_t + D]$に対する加重平均として導出される; Dは経験的に選択される。[8] グローバルアプローチとは異なり、ローカルアライメントベクトルa_tは固定次元、すなわち$\in \mathbb{R}^{2D+1}$となる。モデルの変形として、以下の2つを考える。

単調アライメント(local-m) – ソース配列とターゲット配列がほぼ単調にアライメントされていると仮定して、単純に$p_t = t$とする。アライメントベクトルa_tは式(7)に従って定義される[9]。

予測アライメント(local-p) – 単調アライメントを仮定する代わりに、我々のモデルは以下のようにアライメントされた位置を予測する:

$$p_t = S \cdot \text{sigmoid}(\boldsymbol{v}_p^\top \tanh(\boldsymbol{W}_p \boldsymbol{h}_t)), \quad (9)$$

W_p と v_p は位置を予測するために学習されるモデルパラメータである。S は原文の長さである。シグモイドの結果、$p_t \in [0, S]$となる。p_t 近傍のアライメント点を好むために、p_t を中心とするガウス分布を置く。具体的には、アライメントの重みは次のように定義される:

$$\boldsymbol{a}_t(s) = \text{align}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) \exp\left(-\frac{(s - p_t)^2}{2\sigma^2}\right) \quad (10)$$

式(7)と同じ整列関数を用い、標準偏差は経験的に$\sigma = \frac{D}{2}$とする。p_t は実数であり、s は p_t を中心とする窓の中の整数であることに注意。

[8] ウィンドウが文の境界を越える場合、単に外側の部分を無視し、ウィンドウ内の単語を考慮する。[9] local-mは、ベクトルa_tが固定長で短いことを除けば、グローバルモデルと同じである。

local-p is similar to the local-m model except that we dynamically compute p_t and use a truncated Gaussian distribution to modify the original alignment weights $\text{align}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s)$ as shown in Eq. (10). By utilizing p_t to derive $\boldsymbol{a}_t$, we can compute backprop gradients for $\boldsymbol{W}_p$ and $\boldsymbol{v}_p$. This model is differentiable almost everywhere.

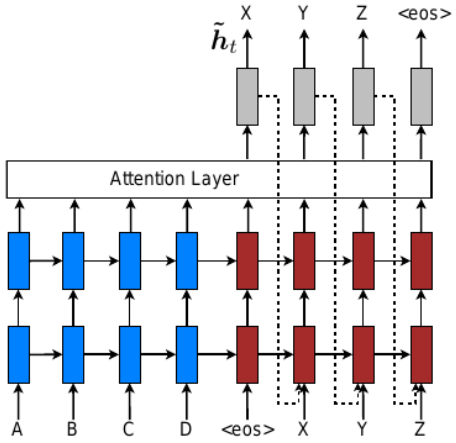

図4: 入力供給アプローチ – 注意ベクトルh_t^*は、過去のアライメント決定についてモデルに情報を与えるために、次の時間ステップの入力として供給される。

(Gregor et al., 2015)との比較 – 画像生成タスクに対して、我々の局所的注意に非常に似た選択的注意メカニズムを提案している。彼らのアプローチにより、モデルは様々な位置とズームの画像パッチを選択することができる。代わりに、すべてのターゲット位置に対して同じ「ズーム」を使用することで、定式化を大幅に簡略化し、依然として良好な性能を達成している。

3.3 入力供給アプローチ

我々の提案するグローバルアプローチとローカルアプローチでは、注意の決定は独立して行われるため、最適とは言えない。一方、標準的なMTでは、どの原語が翻訳されたかを追跡するために、翻訳プロセス中にカバレッジセットが維持されることが多い。同様に、注意NMTでは、過去のアライメント情報を考慮して、アライメントの決定を共同で行うべきである。そこで、図4に示すように、注意ベクトルh_t^*を次の時間ステップの入力と連結する入力フィードアプローチを提案する。[11] このような接続を持つことの効果は2つある:(a)以前のアライメントの選択をモデルに完全に認識させたい、(b)水平方向と垂直方向の両方にまたがる非常に深いネットワークを作成する。

Comparison to other work – Bahdanauら(2015)は、我々の$\tilde{h}_t$と同様に、コンテキストベクトルを用いて、後続の隠れ状態を構築しており、これも「カバレッジ」効果を実現することができる。しかし、このような接続が本研究で行ったような有用かどうかの分析は行われていない。

また、我々のアプローチはより一般的である。図4に示すように、非注意モデルを含む一般的なスタッキングリカレントアーキテクチャに適用することができる。

Xuら(2015)は、キャプション生成プロセス中にモデルが画像の全ての部分に等しく注意を払うようにするために、学習目的に追加の制約を加えた二重注意アプローチを提案する。このような制約は、先に述べたNMTにおけるカバレッジセット効果を捉えるのにも有効である。しかし、入力-フィード・アプローチは、モデルが適切と判断した注意の制約を柔軟に決定できるため、我々は入力-フィード・アプローチを使用することにした。

4 Experiments

英語とドイツ語のWMT翻訳タスクにおいて、我々のモデルの有効性を両方向から評価する。newstest2013 (3000文)を開発セットとして使用し、ハイパーパラメータを選択する。翻訳性能は、newstest2014(2737文)とnewstest2015(2169文)の大文字小文字を区別したBLEU(Papineni et al.)(Luong et al., 2015)に従い、2種類のBLEUを用いた翻訳品質を報告する:(a)既存のNMT作業と比較できるようにトークン化[12] BLEU、(b)WMT結果と比較できるようにNIST[13] BLEU。

4.1 トレーニングの詳細

全てのモデルは4.5M文のペア(116M英単語、110Mドイツ語単語)からなるWMT'14学習データで学習される。(Jean et al., 2015)と同様に、両言語で最も頻度の高い単語の上位50K語を語彙に限定する。これらのショートリストされた語彙にない単語は、ユニバーサルトークンに変換される。

NMTシステムを学習する際、(Bahdanau et al., 2015; Jean et al., 2015)に従い、長さが50語を超える文ペアをフィルタリングし、ミニバッチをシャッフルする。我々のスタッキングLSTMモデルは4層で、それぞれ1000セル、1000次元の埋め込みを持つ。(a) パラメータは [-0.1, 0.1] で一様に初期化、(b) プレーン SGD を用いて 10 エポック学習、

[11] n を LSTM セルの数とすると、最初の LSTM 層の入力サイズは 2n であり、それ以降の層の入力サイズは n である。

[12] All texts are tokenized with `tokenizer.perl` and BLEU scores are computed with `multi-bleu.perl`.

[13] With the `mteval-v13a` script as per WMT guideline.

System	Ppl	BLEU
Winning WMT'14 system – phrase-based + large LM (Buck et al., 2014)		20.7
Existing NMT systems		
RNNsearch (Jean et al., 2015)		16.5
RNNsearch + unk replace (Jean et al., 2015)		19.0
RNNsearch + unk replace + 大規模語彙 + アンサンブル8モデル (Jean et al., 2015)		**21.6**
Our NMT systems		
Base	10.6	11.3
Base + reverse	9.9	12.6 (+1.3)
Base + reverse + dropout	8.1	14.0 (+1.4)
Base + reverse + dropout + global attention (location)	7.3	16.8 (+2.8)
Base + reverse + dropout + global attention (location) + feed input	6.4	18.1 (+1.3)
Base + reverse + dropout + local-p attention (general) + feed input	5.9	19.0 (+0.9)
Base + reverse + dropout + local-p attention (general) + feed input + unk replace		20.9 (+1.9)
Ensemble 8 models + unk replace		**23.0** (+2.1)

表1:WMT'14英独の結果-newstest2014における各システムの当惑度(ppl)とトークン化されたBLEUスコアを示す。最良のシステムを太字で強調し、連続するシステム間で漸進的な改善を斜体で示す。local-pは予測アライメントを持つ局所的な注意を指す。各注意モデルについて、パラレンテーゼで使用されるアライメントスコア関数を示す。

(c) 単純な学習率スケジュールを採用 – 学習率 1 で開始、5 エポック後にエポックごとに学習率を半分にし始め、(d) ミニバッチサイズは 128、(e) 正規化勾配はノルムが 5 を超えると再スケールされる。さらに、確率 0.2 f のドロップアウトも使用する。または、(Zaremba et al., 2015)が提案する我々のLSTM。ドロップアウトモデルについては、12エポック学習し、8エポック後に学習率を半減させ始める。局所注意モデルについては、経験的にウィンドウサイズD = 10とした。

我々のコードはMATLABで実装されている。単一のGPUデバイスTesla K40上で実行した場合、1秒間に1Kの目標単語の速度を達成する。モデルを完全に学習するには7〜10日かかる。

4.2 英語-ドイツ語の結果

英語-ドイツ語タスクにおける我々のNMTシステムを、他の様々なシステムと比較する。WMT'14(Buck他、2014)の優勝システム、言語モデルが膨大な単言語テキストで学習されたフレーズベースのシステム、Common Crawlコーパスなどが含まれる。エンドツーエンドのNMTシステムについては、我々の知る限り、(Jean et al., 2015)がこの言語ペアと現在のSOTAシステムを実験した唯一の研究である。我々はいくつかの注意モデルの結果のみを提示し、残りはセクション5で分析する。

As shown in Table 1, we achieve pro-

(a)原文を反転させた場合の漸進的な改善、+1.3 BLEU、(b)ドロップアウトを使用した場合の漸進的な改善、+1.4 BLEU。その上、(c)グローバル注意アプローチは+2.8 BLEUの大幅なブーストを与え、我々のモデルをBahdanauら(2015)のベース注意システム(行RNNSearch)よりわずかに良くする。(d)入力フィード・アプローチを使用した場合、+1.3 BLEUという別の顕著な利得を捉え、彼らのシステムを上回る。予測アライメントを用いた局所注意モデル(row local-p)はさらに優れており、大域的注意モデルの上に+0.9 BLEUの更なる改善を与える。Luong et al., 2015)で報告された、当惑度が翻訳品質と強く相関する傾向を観察することは興味深い。合計で、我々は非注意ベースラインに対して5.0 BLEUポイントの有意な利得を達成した。これは、ソース反転やドロップアウトなどの既知の技術をすでに含んでいる。

(Luongら, 2015; Jeanら, 2015)で提案された未知の置換技術は、+1.9 BLEUの別の良い利得をもたらし、我々の注意モデルが未知の作品に対して有用なアライメントを学習することを実証する。最後に、様々な設定の8つの異なるモデルをアンサンブルすることで、例えば、異なる注意アプローチ、ドロップアウトの有無などを用いて、我々は23.0 BLEUという新しいSOTA結果を達成し、既存の

best system (Jean et al., 2015) by +1.4 BLEU.

System	BLEU
Top - NMT + 5-gram rerank (Montreal)	24.9
アンサンブル8モデル+unk置換	**25.9**

表2:WMT'15英独結果 - WMT'15で優勝したエントリーと、newstest2015で我々のベストエントリーのNIST BLEUスコア。

WMT'15における最新の結果 - 我々のモデルはWMT'14でわずかに少ないデータで訓練されたにもかかわらず、異なるテストセットにうまく一般化できることを実証するために、newstest2015でテストした。表2に示すように、我々の最良のシステムは25.9 BLEUという新しいSOTA性能を確立し、NMTと5-gram LMリランカーに裏打ちされた既存の最良のシステムを+1.0 BLEU上回った。

4.3 ドイツ語と英語の結果

WMT'15翻訳タスクについて、ドイツ語から英語への同様の実験を行う。我々のシステムはまだSOTAシステムの性能に及ばないが、表3に示すように、BLEUの点で大きく漸進的な利得を持つ我々のアプローチの有効性を示す。注意メカニズムは+2.2BLEUの利得を与え、その上で、入力フィードアプローチから最大+1.0BLEUの別のブーストを得る。より良いアライメント関数を使用すると、コンテンツベースのドットプロダクト1がドロップアウトとともに+2.7BLEUの別の利得を得る。最後に、未知語置換技術を適用した場合、さらに+2.1 BLEUを捉え、希少語の整列における注意の有用性を実証した。

5 Analysis

我々は、学習、長い文章を扱う能力、注意アーキテクチャの選択、アライメント品質の観点から、我々のモデルをより良く理解するために広範な分析を行う。ここで報告する結果はすべて英独newstest2014でのものである。

5.1 学習曲線

表1に示すように、互いに重ね合わされたモデルを比較する。図5では、非注意モデルと注意モデルが明確に分かれていることが観察される。入力供給アプローチと局所注意モデルは、テストコストを低くする能力も実証している。ドロップアウトを伴う非注意モデル(青)は

System	Ppl.	BLEU
WMT'15 systems		
SOTA - フレーズベース(エディンバーグ)		**29.2**
NMT+5-gram再ランク(MILA)		27.6
Our NMT systems		
Base (reverse)	14.3	16.9
+ global (location)	12.7	19.1 (+2.2)
+ global (location) + feed	10.9	20.1 (+1.0)
+ global (dot) + drop + feed	9.7	22.8 (+2.7)
+ グローバル(ドット)+ドロップ+フィード+アンク		24.9 (+2.1)

表3:WMT'15独英の結果-各システムの性能(表1と同様)。ベースシステムはすでにソース反転を含んでおり、その上にグローバルアテンション、ドロップアウト、入力フィード、unk置換を追加している。

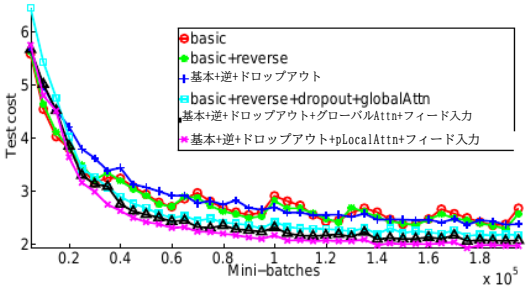

図5:学習曲線 - newstest2014における英独NMTのテストコスト(ln perplexity)。

+ curve)は他の非ドロップアウトモデルよりも学習速度が遅いが、時間が経つにつれて、テスト誤差を最小化する点でよりロバストになる。

5.2 長い文章の翻訳による効果

我々は(Bahdanau et al., 2015)に従い、同様の長さの文をグループ化し、グループごとにBLEUスコアを計算する。図6は、我々の注意モデルが、長い文を扱う上で、非注意モデルよりも効果的であることを示している:文が長くなるにつれて品質が低下することはない。我々の最良のモデル(青+曲線)は、すべての長さのバケットにおいて、他のすべてのシステムを凌駕している。

5.3 注意アーキテクチャの選択

我々はセクション3で説明したように、異なる注意モデル(global、local-m、local-p)と異なるアライメント関数(location、dot、general、concat)を検証する。リソースが限られているため、全ての可能な組み合わせを実行することはできない。しかし、表4の結果は、異なる選択肢についてある程度の示唆を与えてくれる。

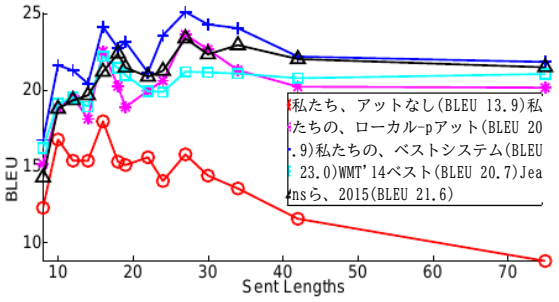

図6: 長さ分析 - 文が長くなるにつれて、異なるシステムの翻訳品質が変化する。

System	Ppl	BLEU	
		Before	After unk
global (location)	6.4	18.1	19.3 (+1.2)
global (dot)	6.1	18.6	20.5 (+1.9)
global (general)	6.1	17.3	19.1 (+1.8)
local-m (dot)	>7.0	x	x
local-m (general)	6.2	18.6	20.4 (+1.8)
local-p (dot)	6.6	18.0	19.6 (+1.9)
local-p (general)	**5.9**	**19**	**20.9 (+1.9)**

表4:注意アーキテクチャ - 異なる注意モデルのパフォーマンス。2つのローカル-m(ドット)モデルを学習した。

位置情報ベースの関数は、良いアライメントを学習しない:グローバル(位置)モデルは、他のアライメント関数を使用する場合と比較して、未知の単語置換を実行する場合にのみ、わずかな利得を得ることができる。[14] コンテンツベースの関数については、我々の実装concatは良い性能をもたらさないので、その理由を理解するためにより多くの分析を行う必要がある。[15] ドットがグローバル注意にうまく機能し、ローカル注意に一般的であることが観察するのは興味深い。異なるモデルの中で、予測アライメント(localp)を持つ局所的注意モデルは、当惑度とBLEUの両面で最良である。

5.4 アライメント品質

注意モデルの副産物として、単語の整列がある。一方、(Bahdanau et al., 2015)は可視化された

[14] アライメント関数の違いによるアライメントの取得方法に微妙な違いがある。y_{t-1}を入力とし、y_tを予測する前にh_t, a_t, c_t, h_t^*を計算する時間ステップtにおいて、(a)位置情報ベースのアライメント関数では予測単語y_t、(b)内容情報ベース関数では入力単語y_{t-1}のアライメント重みとしてアライメントベクトルa_tを用いる。

[15] concatを用いると、異なるモデルで達成される当惑度は6.7(global)、7.1(local-m)、7.1(local-p)となる。このような高い当惑度は、行列W_aを簡略化して、h_sに対応する部分を恒等式に設定したためと考えられる。

Method	AER
global (location)	0.39
local-m (general)	0.34
local-p (general)	0.36
ensemble	0.34
Berkeley Aligner	0.32

表6:AERスコア-RWTH英独アライメントデータに対する各種モデルの結果。

いくつかのサンプル文のアライメントと、作業注意モデルの指標としての翻訳品質の向上が観察されたが、全体として学習されたアライメントを評価した研究はない。一方、アライメントエラーレート(AER)指標を用いてアライメント品質を評価することにした。

508の英独Europarl文に対してRWTHが提供するゴールドアライメントデータが与えられたとき、我々は参照に一致する翻訳を生成するために注意モデルを「強制」デコードする。対象単語ごとに最も高いアライメント重みを持つ原語を選択することで、1対1のアライメントのみを抽出する。それにもかかわらず、表6に示すように、Berkeley aligner (Liang et al., 2006)によって得られた1対多のアライメントに匹敵するAERスコアを達成することができた[16]。

また、局所的な注意モデルによって生成されたアライメントは、大域的な注意モデルよりも低いAERを達成することがわかった。アンサンブルによって得られたAERは、良好ではあるが、ローカルなAERよりも優れておらず、AERと翻訳スコアはよく相関していないというよく知られた観察結果を示唆している(Fraser and Marcu, 2007)。付録Aにアライメントの可視化を示す。

5.5 サンプル翻訳

表5に、両方向の平行移動の例を示す。Miranda Kerr ”や ”Roger Dow ”などの名前を正しく翻訳する際の注意モデルの効果を観察することは魅力的である。非注意モデルは、言語モデルの観点からは賢明な名前を生成するが、正しい翻訳を行うためにソース側からの直接的な接続を欠いている。また、2番目の例では、二重否定されたフレーズである「相容れないでない」を翻訳する必要があるという興味深いケースも観察された。注意モデルは「nicht.. unvereinbar」を正しく生成するのに対し、非注意モデルは「nicht vereinbar」を生成し、

[16] WMT の 508 文ペアと 1M 文ペアを連結し、Berkeley aligner を実行する。

英語-ドイツ語翻訳	
オーランド・ブルームとミランダ・カーは、いまだにお互いを愛している。	
ref	Orlando Bloom und Miranda Kerr lieben sich noch immer
best	Orlando Bloom und Miranda Kerr lieben einander noch immer .
base	Orlando Bloom und **Lucas Miranda** lieben einander noch immer .
src	″ We ' re pleased the FAA recognizes that an enjoyable passenger experience is not incompatible with safety and security , ″ said Roger Dow , CEO of the U.S. Travel Association .
ref	" Wir freuen uns , dass die FAA erkennt , dass ein angenehmes Passagiererlebnis nicht im Widerspruch zur Sicherheit steht " , sagte Roger Dow , CEO der U.S. Travel Association .
best	″ Wir freuen uns , dass die FAA anerkennt , dass ein angenehmes ist nicht mit Sicherheit und Sicherheit unvereinbar ist ″ , sagte Roger Dow , CEO der US - die .
base	″ Wir freuen uns über die <unk> , dass ein <unk> <unk> mit Sicherheit nicht **vereinbar** ist mit Sicherheit und Sicherheit ″ , sagte Roger **Cameron** , CEO der US - <unk> .
ドイツ語-英語翻訳	
src	In einem Interview sagte Bloom jedoch , dass er und Kerr sich noch immer lieben .
ref	However , in an interview , Bloom has said that he and Kerr still love each other .
best	In an interview , however , Bloom said that he and Kerr still love .
base	However , in an interview , Bloom said that he and **Tina** were still <unk> .
src	Wegen der von Berlin und der Europäischen Zentralbank verhängten strengen Sparpolitik in Verbindung mit der Zwangsjacke , in die die jeweilige nationale Wirtschaft durch das Festhalten an der gemeinsamen Währung genötigt wird , sind viele Menschen der Ansicht , das Projekt Europa sei zu weit gegangen
ref	ベルリンと欧州中央銀行が課した緊縮財政は、共通通貨の遵守を通じて国民経済に課された拘束力と相まって、多くの人がプロジェクトヨーロッパが行き過ぎたと考えるようになった。
best	Because of the strict austerity measures imposed by Berlin and the European Central Bank in connection with the straitjacket in which the respective national economy is forced to adhere to the common currency , many people believe that the European project has gone too far .
base	欧州中央銀行と連邦中央銀行が、単一通貨を前にして国民経済に厳しい緊縮財政を課していることにプレッシャーをかけているため、多くの人々は、欧州プロジェクトは行き過ぎたものだと考えている。

表5:翻訳例-各例について、ソース(src)、人間の翻訳(ref)、我々の最良のモデルからの翻訳(best)、非注意モデルの翻訳(base)を示す。正しい翻訳セグメントをイタリック体で示し、間違った翻訳セグメントを太字でいくつか強調する。

「互換性がない」ことを意味する[17]。注意モデルはまた、最後の例のように長い文章を翻訳する際にも優れていることを示している。

6 Conclusion

本論文では、ニューラル機械翻訳のための2つのシンプルで効果的な注意メカニズムを提案する:常にすべてのソース位置を見るグローバルアプローチと、一度にソース位置のサブセットにのみ注意を向けるローカルアプローチである。英語とドイツ語のWMT翻訳タスクにおいて、我々のモデルの有効性を両方向に検証する。我々の局所的な注意は、ドロップアウトのような既知の技術を既に取り入れた非注意モデルに対して、最大5.0 BLEUの大きな利得をもたらす。

英語からドイツ語への翻訳方向において、我々のアンサンブルモデルはWMT'14とWMT'15の両方で新しい最先端の結果を確立し、NMTモデルとn-gram LMリランカーに裏打ちされた既存の最良システムを1.0 BLEU以上上回った。
様々なアライメント関数を比較し、どの関数がどの注意モデルに最適であるかを明らかにした。我々の分析によると、注意に基づくNMTモデルは、例えば名前の翻訳や長い文章の処理など、多くの場合、非注意のものよりも優れている。

Acknowledgment

Bloomberg L.P.からの寄贈に感謝する。また、NVIDIA Corporation の Tesla K40 GPU の寄贈による支援に感謝する。

[17] この文献では、「im Widerspruch zu etwas stehen」という、より派手な「非互換性」の訳語を使用している。しかし、どちらのモデルも「受動的な経験」を翻訳することはできなかった。

Andrew Ngと彼のグループ、そして彼らの計算資源を使わせてくれたStanford Research Computingに感謝する。Russell Stewart には、モデルについて有益な議論をいただいた。最後に、Quoc Le、Ilya Sutskever、Oriol Vinyals、Richard Socher、Michael Kayser、Jiwei Li、Panupong Pasupat、Kelvin Guu、Stanford NLP Groupのメンバー、および貴重なコメントとフィードバックをくれた匿名査読者に感謝する。

References

[Bahdanau et al.2015] D. Bahdanau, K. Cho, and Y. Bengio. 2015. Neural machine translation by jointly learning to align and translate. In ICLR.

[Buck et al.2014] Christian Buck, Kenneth Heafield, and Bas van Ooyen. 2014. N-gram counts and language models from the common crawl. In LREC.

[Cho et al.2014] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In EMNLP.

[Fraser and Marcu2007] Alexander Fraser and Daniel Marcu. 2007. Measuring word alignment quality for statistical machine translation. Computational Linguistics, 33(3):293–303.

[Gregor et al.2015] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. 2015. DRAW: A recurrent neural network for image generation. In ICML.

[Jean et al.2015] Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. On using very large target vocabulary for neural machine translation. In ACL.

[Kalchbrenner and Blunsom2013] N. Kalchbrenner and P. Blunsom. 2013. Recurrent continuous translation models. In EMNLP.

[Koehn et al.2003] Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In NAACL.

[Liang et al.2006] P. Liang, B. Taskar, and D. Klein. 2006. Alignment by agreement. In NAACL.

[Luong et al.2015] M.-T. Luong, I. Sutskever, Q. V. Le, O. Vinyals, and W. Zaremba. 2015. Addressing the rare word problem in neural machine translation. In ACL.

[Mnih et al.2014] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. 2014. Recurrent models of visual attention. In NIPS.

[Papineni et al.2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In ACL.

[Sutskever et al.2014] I. Sutskever, O. Vinyals, and Q. V. Le. 2014. Sequence to sequence learning with neural networks. In NIPS.

[Xu et al.2015] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In ICML.

[Zaremba et al.2015] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2015. Recurrent neural network regularization. In ICLR.

アライメントの可視化

図7に、我々の異なる注意モデルによって生成されたアライメント重みを可視化する。局所的注意モデルの可視化は、大域的注意モデルの可視化よりもはるかにシャープである。この対比は、局所的な注意は毎回単語のサブセットにのみ焦点を当てるように設計されているという我々の予想と一致する。また、英語からドイツ語に翻訳し、原文の英文を逆にするので、白いストライドは ”reality ”と ””である。グローバルアテンションモデルでは、興味深いアクセスパターンが明らかになった。それは、ソースシーケンスの開始を指し示す傾向がある。

(Bahdanau et al., 2015)のアライメント可視化と比較すると、我々のアライメントパターンは彼らのものほどシャープではない。このような違いは、(Bahdanau et al., 2015)で行われたように、英語からドイツ語への翻訳がフランス語への翻訳よりも難しいという事実に起因している可能性があり、これは今後の研究で検討すべき興味深い点である。

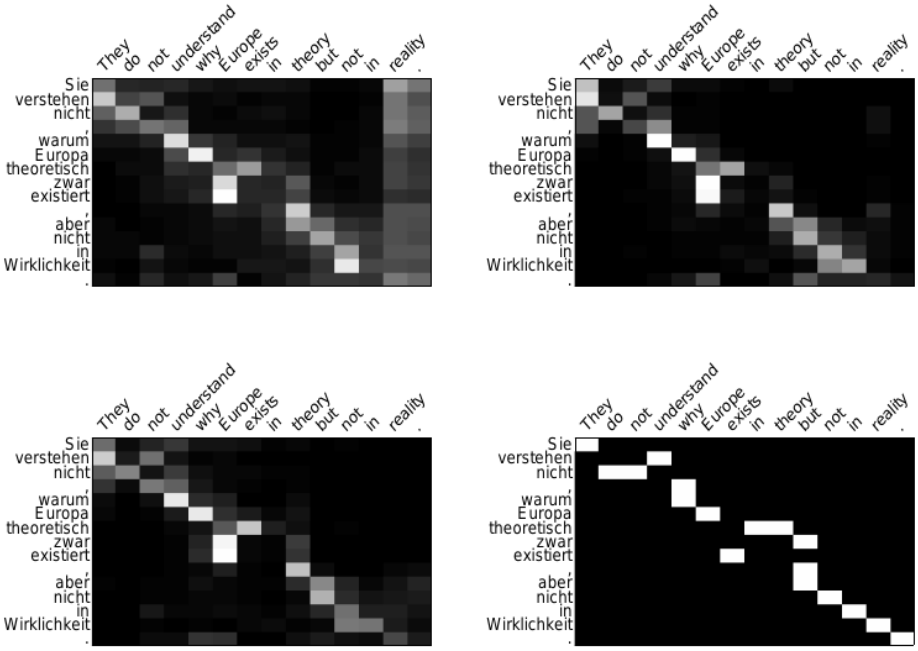

図7:アライメントの可視化 – 様々なモデルによって学習された注目重みの画像を示す:(左上)グローバル、(右上)ローカル-m、(左下)ローカル-p。右下に金のアラインメントが表示されている。